Barbara Marques | Easley Cohort

May 11, 2021

Predicting Programming Language of GitHub Cybersecurity Repos

Natural Language Processing

Our Agenda

1 Executive Summary

2 Acquisition & Exploration

3 Modeling

4 Summary of Findings

Executive Summary

Goals

- Build a model to predict the primary programming language used in a GitHub repository
- Identify primary programming languages used by in cybersecurity repositories

Process

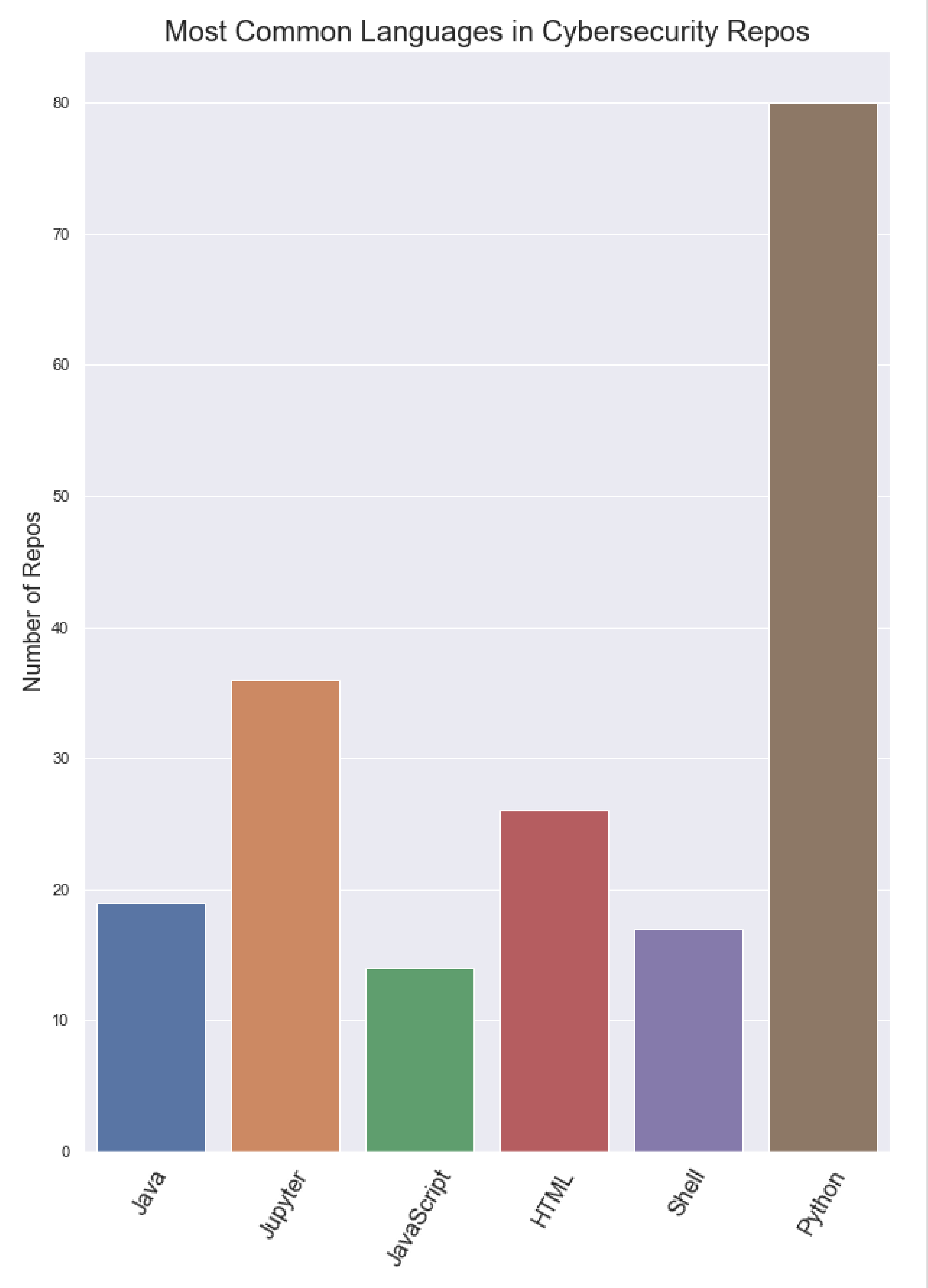
- Readme files were scraped from GitHub repositories
- Cleaned and prepared for NLP modeling
- Modeled on three different classification models

Findings

- Majority used Python, HTML or Jupyter Notebook as the primary language
- K-Nearest Neighbors model out-performed baseline with an accuracy of 49%

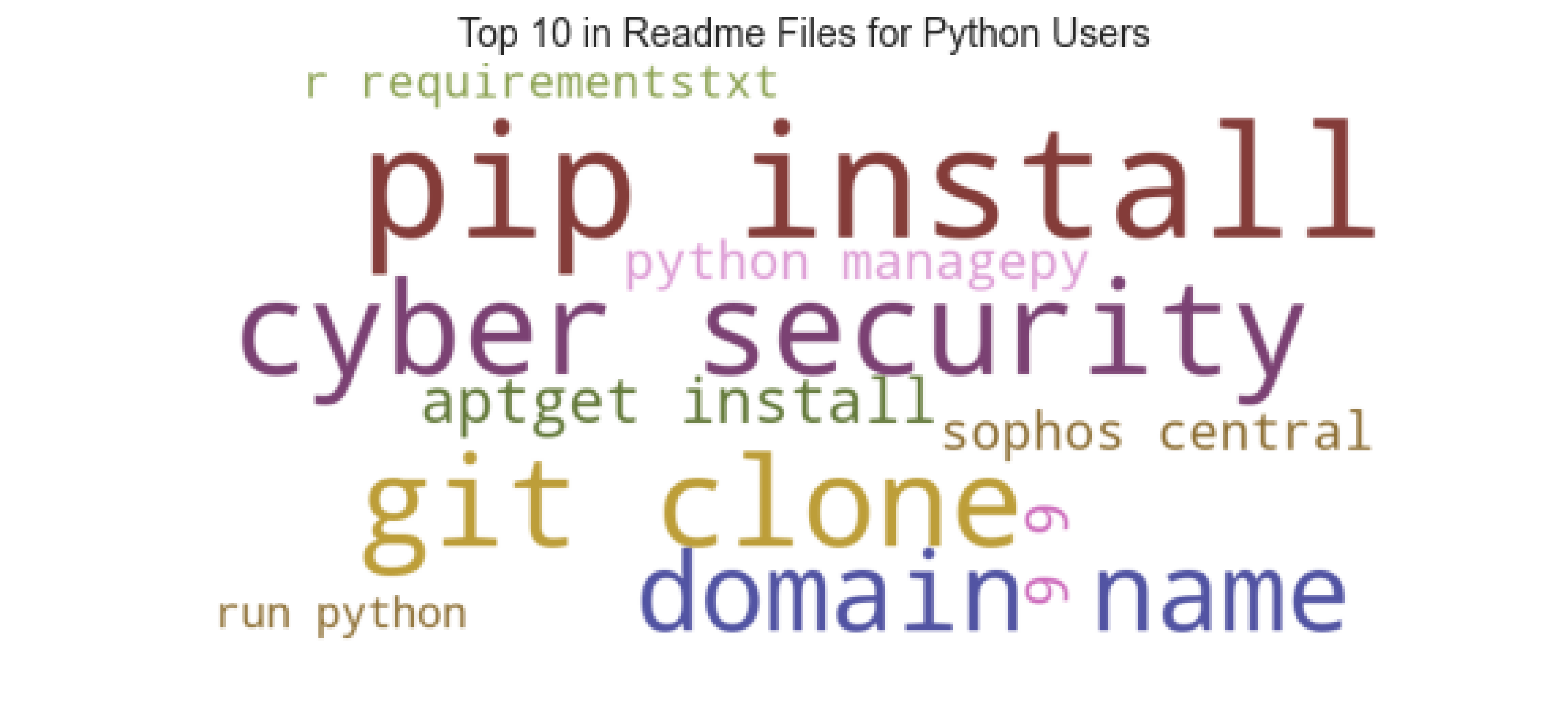
What languages are most common in cybersecurity GitHub repositories?

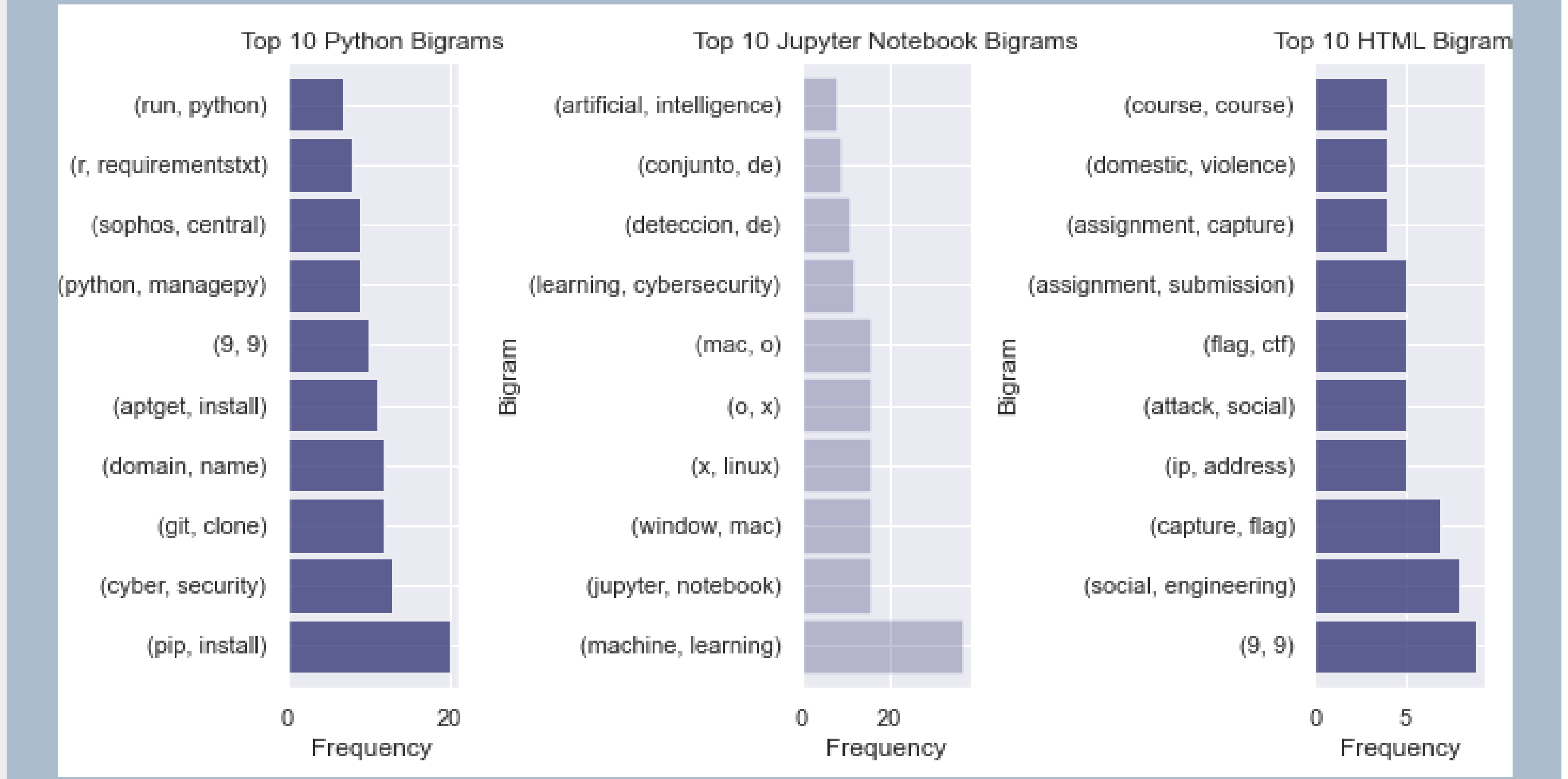
- 342 Readme Files
- 30 languages
- Model based on Top 6
- Python is most common



Acquiring, Preparing & Exploring the Data

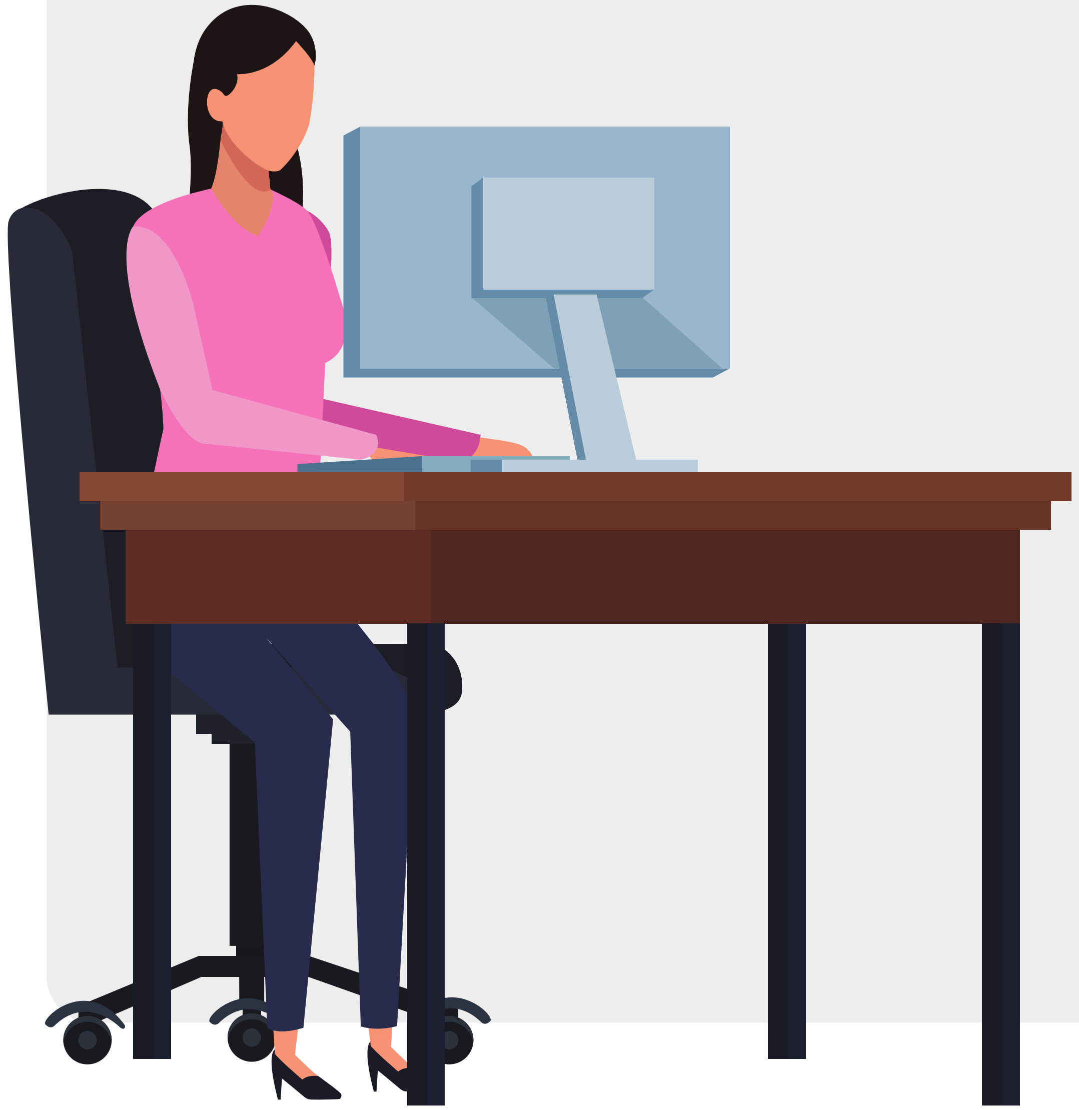
- Scraped 470 Readme files
- Top 6 languages kept
- 240 readme files were analyzed,
- Created n-grams using NLTK
- Word clouds/bar plots to visualize word importance
- Data was split and vectorized in preparation for modeling

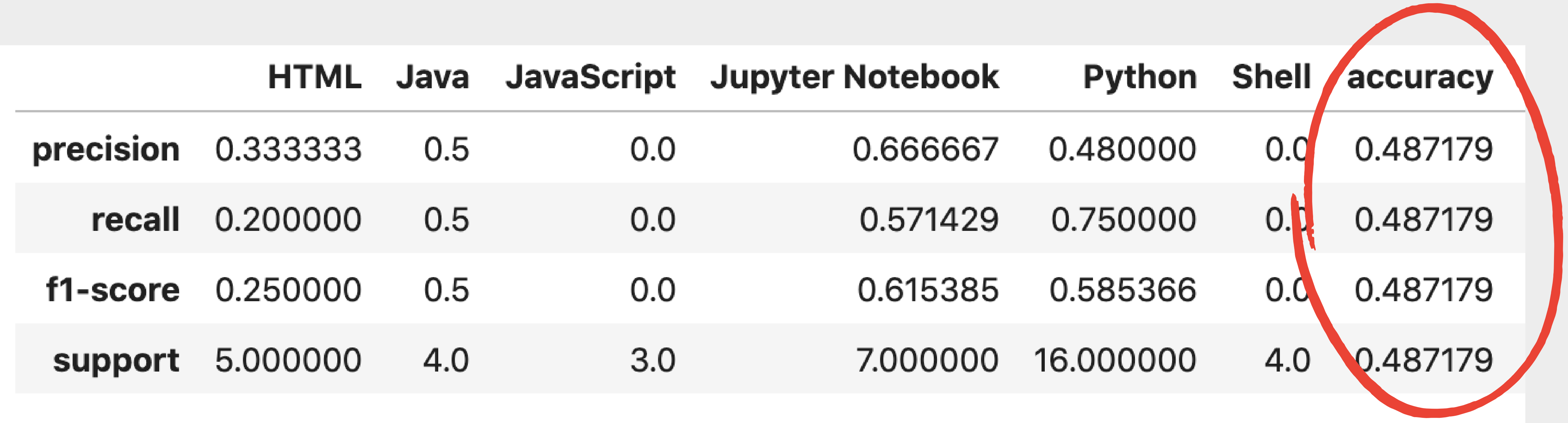

Top 10 in Readme Files for Python Users



Modeling

- Baseline Accuracy: of 42%
- Modeled sample on: logistic regression, random forest classifier and KNN.
- Using a K-Nearest Neighbor model on test data, I was able to predict the primary language with an accuracy of 49%.



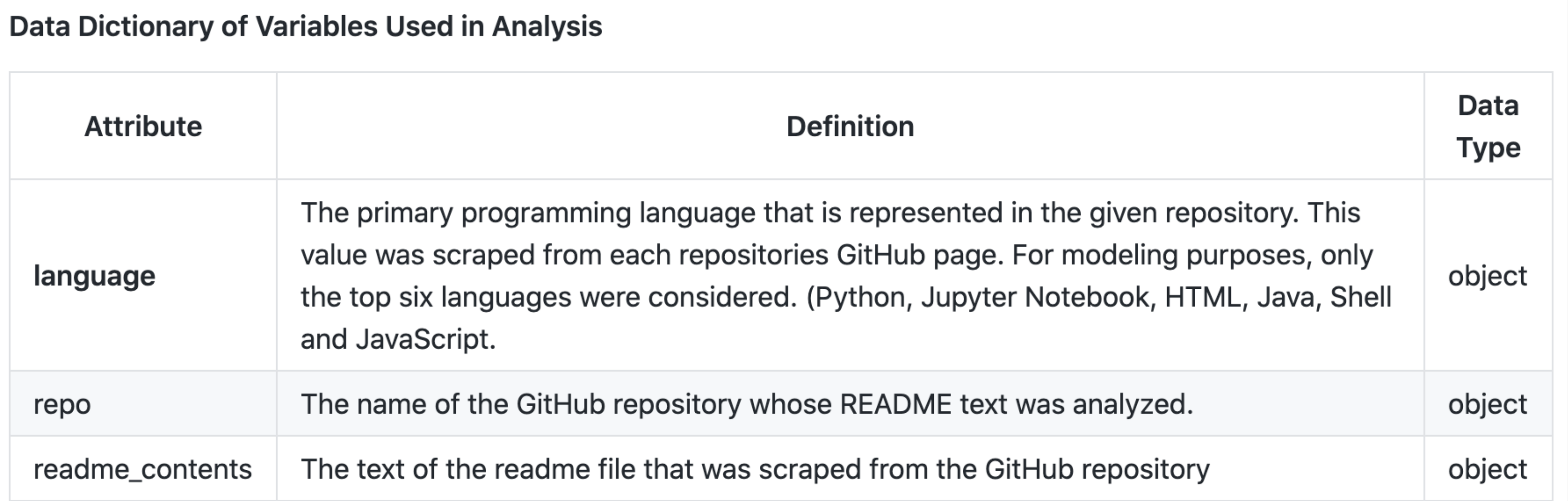


	HTML	Java	JavaScript	Jupyter Notebook	Python	Shell	accuracy
precision	0.333333	0.5	0.0	0.666667	0.480000	0.0	0.487179
recall	0.200000	0.5	0.0	0.571429	0.750000	0.0	0.487179
f1-score	0.250000	0.5	0.0	0.615385	0.585366	0.0	0.487179
support	5.000000	4.0	3.0	7.000000	16.000000	4.0	0.487179

Conclusion

- 240 cybersecurity README files analyzed
- Python was the predominant language
- K-Nearest Neighbor model predicted the primary language of cybersecurity repositories with an accuracy of 49%.
- This beats baseline performance of 42%.

Appendix



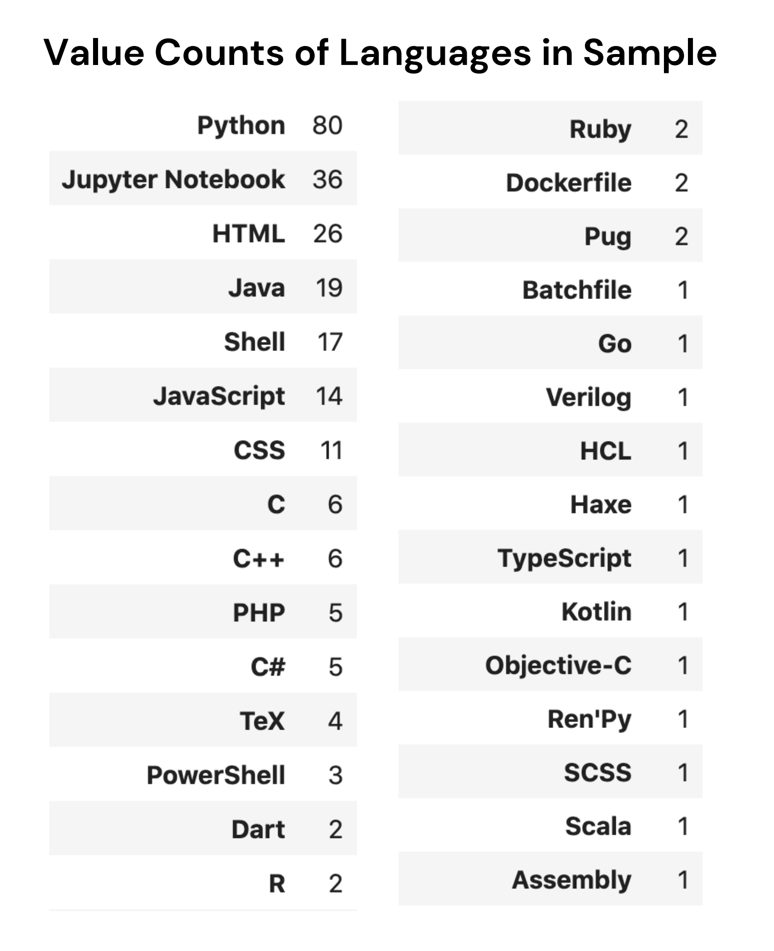
Data Dictionary of Variables Used in Analysis

Attribute	Definition	Data Type
language	The primary programming language that is represented in the given repository. This value was scraped from each repositories GitHub page. For modeling purposes, only the top six languages were considered. (Python, Jupyter Notebook, HTML, Java, Shell and JavaScript.	object
repo	The name of the GitHub repository whose README text was analyzed.	object
readme_contents	The text of the readme file that was scraped from the GitHub repository	object

K-Nearest Neighbor Model:

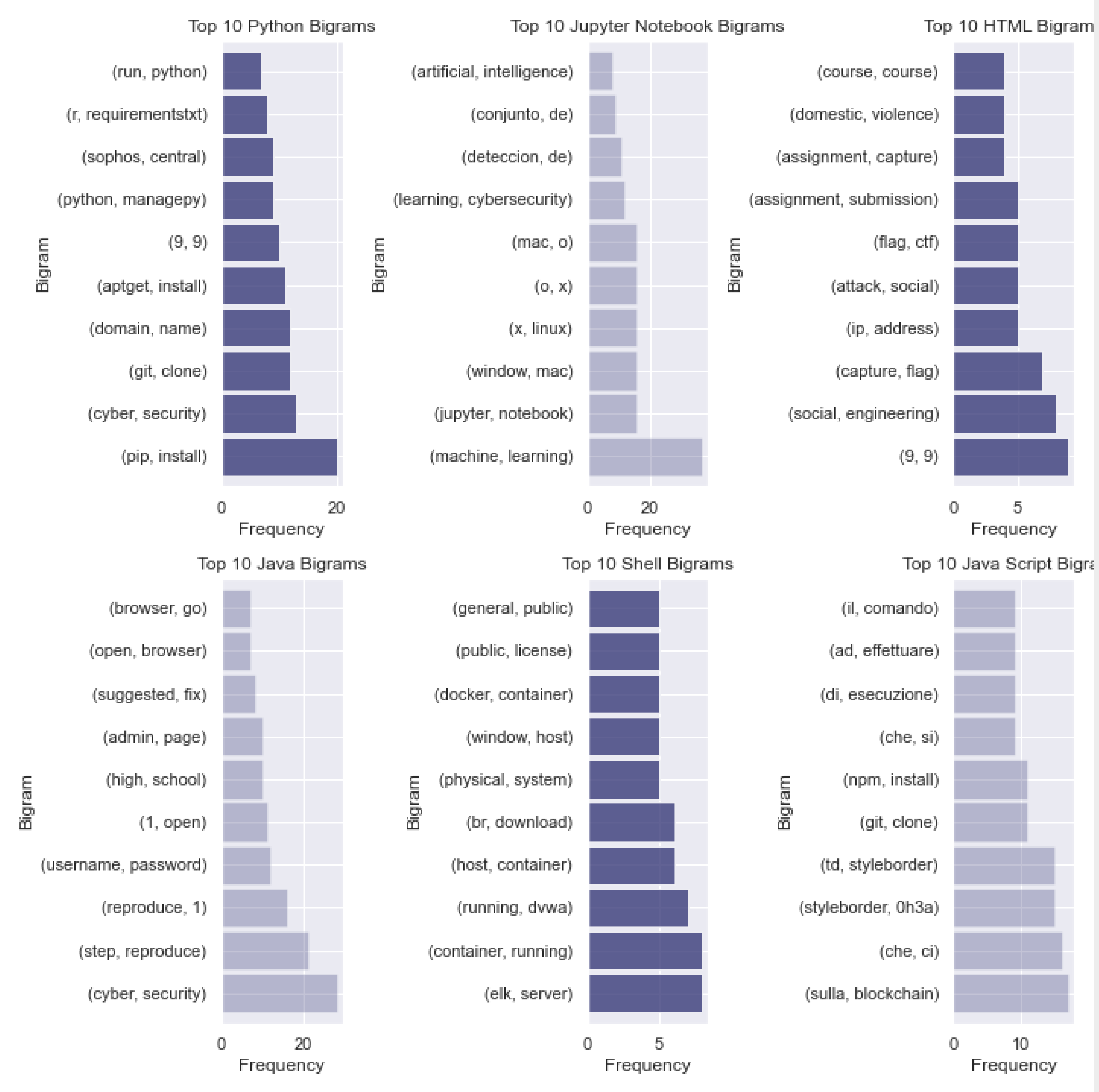
- K-Nearest Neighbor on Test: Accuracy of 49%
- Baseline Accuracy: 42%
- Train Accuracy: 63%
- Validate Accuracy: 43%

For additional information, please see the README.md file @ https://github.com/barbmarques/individual-nlp-project/blob/main/README.md



Value Counts of Languages in Sample

Python	80	**Ruby**	2
Jupyter Notebook	36	**Dockerfile**	2
HTML	26	**Pug**	2
Java	19	**Batchfile**	1
Shell	17	**Go**	1
JavaScript	14	**Verilog**	1
CSS	11	**HCL**	1
C	6	**Haxe**	1
C++	6	**TypeScript**	1
PHP	5	**Kotlin**	1
C#	5	**Objective-C**	1
TeX	4	**Ren'Py**	1
PowerShell	3	**SCSS**	1
Dart	2	**Scala**	1
R	2	**Assembly**	1


Top 10 Python Bigrams
(run, python)
(r, requirementstxt)
(sophos, central)
(python, managepy)
(9, 9)
(aptget, install)
(domain, name)
(git, clone)
(cyber, security)
(pip, install)
Bigram
Frequency
Top 10 Jupyter Notebook Bigrams
(artificial, intelligence)
(conjunto, de)
(deteccion, de)
(learning, cybersecurity)
(mac, o)
(o, x)
(x, linux)
(window, mac)
(jupyter, notebook)
(machine, learning)
Bigram
Frequency
Top 10 HTML Bigram
(course, course)
(domestic, violence)
(assignment, capture)
(assignment, submission)
(flag, ctf)
(attack, social)
(ip, address)
(capture, flag)
(social, engineering)
(9, 9)
Bigram
Frequency
Top 10 Java Bigrams
(browser, go)
(open, browser)
(suggested, fix)
(admin, page)
(high, school)
(1, open)
(username, password)
(reproduce, 1)
(step, reproduce)
(cyber, security)
Bigram
Frequency
Top 10 Shell Bigrams
(general, public)
(public, license)
(docker, container)
(window, host)
(physical, system)
(br, download)
(host, container)
(running, dvwa)
(container, running)
(elk, server)
Bigram
Frequency
Top 10 Java Script Bigra
(il, comando)
(ad, effettuare)
(di, esecuzione)
(che, si)
(npm, install)
(git, clone)
(td, styleborder)
(styleborder, 0h3a)
(che, ci)
(sulla, blockchain)
Bigram
Frequency